Proceedings Accuracy 2000, Amsterdam, July 2000

Assessing the Accuracy of Large-Area Land Cover Maps: Experiences from the Multi-Resolution Land-Cover Characteristics (MRLC) Project

Stephen V. Stehman[1], James D. Wickham[2], Limin Yang[3], and Jonathan H. Smith[2]
[1]SUNY College of Environmental Science and Forestry, 320 Bray Hall, Syracuse, NY 13210 USA
Ph. 315 470-6692; Fax. 315 470-6692
Email: svstehma@mailbox.syr.edu
[2]U.S. Environmental Protection Agency, Research Triangle Park, NC 27709 USA
[3]U.S. Geological Survey, EROS Data Center, Sioux Falls, SD 57198 USA

Keywords: remote sensing, sampling, validation

Abstract
Large-area land-cover maps create unique challenges for thematic map accuracy assessment. We describe desirable characteristics for the design of large-area assessments, and propose two-stage cluster sampling as a general framework possessing the flexibility needed to achieve these characteristics. The sampling theory supporting use of the two-stage design is briefly described. We identify a need for further research to evaluate specific design options within this general sampling framework, and we describe some practical and philosophical problems that must still be resolved to enhance the effectiveness of accuracy assessments.

1. Introduction
The Multi-Resolution Land Characteristics (MRLC) mapping program uses 30-meter resolution Landsat Thematic Mapper imagery as the baseline data for a land-cover map of the conterminous United States (*Vogelmann et al., 1998*). The classification scheme consists of 21 classes reflecting an approximate *Anderson (et al., 1976)* Level II detail. Mapping is conducted on a regional basis using ten regions to partition the United States. Accuracy assessment of each geographic region follows completion of the map for that region. The land-cover map assessed has not been aggregated to a minimum mapping unit, although it is anticipated users of the map will often impose such an aggregation. The reference data used to assess accuracy are derived from aerial photography, and the assessment unit is a 30-meter pixel. *Zhu et al. (in press)* and *Yang et al. (2000)* describe additional details of the accuracy assessment protocol.

2. Design Criteria for MRLC Accuracy Assessment
Planning a sampling design for accuracy assessment requires consideration of the assessment objectives and practical constraints. Desirable characteristics for the MRLC accuracy assessment include: 1) the sampling design should satisfy probability sampling protocol; 2) estimates of accuracy parameters should have acceptable precision; 3) a simple design is preferable for proper implementation and data analysis; and 4) costs should be as low as possible.

Probability sampling is defined as a protocol in which the inclusion probabilities for all elements of the sample are known, and inclusion probabilities for all elements of the population are non-zero (*Stehman and Czaplewski, 1998*). Inclusion probabilities are a characteristic of a sampling design, and represent the probability of a population element being included in the sample. Probability sampling is crucial to the scientific defensibility of accuracy assessment. Probability sampling has a long history of use in agriculture, health, and business surveys (*Bellhouse, 1988*), and it is a key element of many natural resource sampling protocols in the United States such as the National Resources Inventory (*Nusser and Goebel, 1997*) and the Forest Inventory and Analysis (*USFS, 1992*). *Kish (1987, p. 23)* notes that probability sampling is the only "feasible method recognized by survey samplers in most practical situations" to ensure a representative sample. Further, the randomization mechanism incorporated in any probability sampling design prevents subjective biases from influencing the sample selected. Recent evidence indicates that convenience sampling can give misleading results in environmental assessments (*Peterson et al., 1999*), so relying on such non-probability samples to provide valid inference in accuracy assessment is a risky proposition. If accuracy assessment is to achieve scientific credibility, probability sampling becomes a necessity.

Within the probability sampling framework, the MRLC sampling design is constructed to provide adequate precision for estimating land-cover class specific estimates for each geographic region, while still maintaining costs within the available budget. Within each region, stratification by mapped land-cover class is implemented to enhance precision of the class-specific estimates (user's accuracies). Sample sizes planned for each land-cover stratum are chosen to accommodate subregional estimates anticipated of interest to users (e.g., state-level estimates). To reduce costs of obtaining reference data, the number of air photos required is limited by implementing a cluster sampling design. The specific characteristics of the clusters defined may vary depending on the geographic region. Simplicity is achieved in part by maintaining equal inclusion probabilities for pixels within a land-cover class stratum within each geographic region. Inclusion probabilities for different land-cover strata differ within a geographic region, and inclusion probabilities for the same land-cover class differ among regions. Because each of the ten geographic regions is treated as a stratum, it is not necessary for every region to be sampled exactly the same. The primary requirements for uniformity among geographic regions are the reference data collection protocol and the definition of agreement used to compare the map and reference classifications.

3. Sampling Design

The general sampling structure implemented for the MRLC accuracy assessment incorporates both cluster sampling and stratification. The first-stage sampling employs a large primary sampling unit (PSU) to restrict the spatial distribution of the sample. The PSU defined for the first geographic regions completed was based on characteristics of the aerial photography used for the reference data. For convenience and to alleviate confusion about the exact nature of the PSU defined from the aerial photography, we have subsequently adopted a 6 *km* by 6 *km* PSU. Once the first-stage sample of PSUs has been selected, these PSUs are then subsampled to obtain the sample pixels.

This general two-stage design structure has great flexibility making it highly advantageous for large-area accuracy assessments. Numerous options are available for sampling at the first and second stages. For example, in one geographic region we selected an equal probability sample of

pixels within each PSU because our objective was to implement a self-weighting design permitting easy analysis (*Zhu et al., in press*). For other applications, it may be desirable to stratify by mapped land-cover class if the objectives specify precise estimates for each class, or to stratify by geographic features such as accessibility or public versus private land to improve cost-effectiveness of the design. In the MRLC design, stratification by mapped land-cover class is a feature common to each geographic region. Once the first-stage sample of PSUs has been selected, the pixels within those PSUs are stratified according to mapped land-cover class. A simple random sample is then obtained from the pooled collection of pixels for each class. This design is analogous to double (or two-phase) sampling for stratification (*Cochran, 1977, Sec. 12.2*), but the cluster sampling imposed at the first-stage complicates the design slightly from the simpler textbook illustrations usually presented. Conditional on the first-stage sample of PSUs, each pixel within a land-cover class stratum has an equal probability of selection.

The current MRLC design protocol does not control the distribution of second-stage sample pixels among the first-stage PSUs. For a given land-cover class, the distribution of the second-stage sample pixels to PSUs will be proportional to that class's representation in the PSUs. That is, more of the second-stage sample pixels will be found in those PSUs for which the class is prevalent. Table 1 illustrates the distribution of pixels to PSUs for the upper Midwest MRLC geographic region when the target sample size is 100 per land-cover class, and the first-stage sampling intensity is 2% of the area. For the more common classes (e.g., 11=water, 41=deciduous forest, 81=pasture/hay, 82=row crops, and 91=forested wetlands), the sampled pixels are generally spread among a large number of PSUs. For example, deciduous forest (class 41) accounts for 19.17% of the population. The sample of 100 deciduous forest sample pixels is distributed such that 77 PSUs have a single sample pixel, 10 PSUs have two sample pixels, and 1 PSU has three sample pixels. Sample pixels of rarer classes (e.g., 21=low density residential, 22=high density residential, 23=commercial or transportation, 83=small grains) may cluster in only a few PSUs. For example, for class 32 (quarries/strip mines, 0.11% of the population), 9 PSUs have a single sample pixel of class 32, 1 PSU has two sample pixels, and 4 PSUs have three sample pixels. The remaining 77 sample pixels of class 32 are found in only 3 other PSUs.

Allocation of the second-stage sample to PSUs is a precision, not a bias issue. In general, a positive within-PSU (i.e., intracluster) correlation for classification error is anticipated reflecting the common phenomenon that classification error tends to be positively spatially autocorrelated. Reducing the effect of positive spatial autocorrelation improves precision, and this is achieved by spreading the sample pixels among more PSUs. To enhance precision for estimated accuracy of rare classes, it may be better to disperse the second-stage sample among more PSUs.

Two alternative sampling protocols to distribute the second-stage sample among more PSUs are described. Both protocols retain stratification by land-cover class. The first alternative is to sample a fixed number of pixels from each PSU. Maximum dispersion among PSUs is achieved by selecting one pixel per PSU. For the first-stage sample shown in Table 1, it would be possible to obtain a sample of 100 pixels from 100 different PSUs in all but three of the land-cover classes (class 31, commercial/transportation; class 32, quarries/strip mines; and class 51, shrubland). Class 32 just misses having 100 PSUs available from the first-stage. A minimum of two pixels per PSU must be sampled to estimate the within PSU component of variation for standard error calculations. This design requires more complex analyses because, within a land-cover stratum, pixels now have unequal inclusion probabilities. As long as the inclusion probabilities are

known, the design still satisfies the probability sampling criterion, but the analysis must account for these unequal inclusion probabilities via proper weighting of the data within strata. Standard stratified random sampling formulas for estimating parameters of the error matrix no longer apply to this version of two-stage sampling. The question arises of whether the potential gain in precision resulting from this design is sufficient to overcome the increased complexity of the analysis.

	Number of PSUs in which the sample size in that PSU equals the column label												
Land Cover Class	1	2	3	4	5	6	7	8	9	>10	Number of PSUs in which at least one sample pixel occurs	Number of first-stage PSUs containing at least one pixel in class	% of popn. in class
11	42	19	4	2	0	0	0	0	0	0	67	529	15.06
21	27	13	1	2	3	2	0	0	1	0	49	340	1.25
22	13	6	1	5	0	1	2	0	0	2	30	231	0.53
23	41	8	9	1	1	0	1	0	0	0	61	394	0.75
31	11	3	1	2	1	0	0	1	1	2	22	59	0.02
32	9	1	4	0	0	0	0	0	0	3	17	99	0.11
33	18	5	8	5	2	3	0	0	0	0	41	160	0.19
41	77	10	1	0	0	0	0	0	0	0	88	497	19.17
42	37	21	5	0	0	1	0	0	0	0	64	459	2.65
43	41	18	3	2	0	1	0	0	0	0	65	406	2.53
51	3	7	4	3	2	3	1	0	0	2	25	70	0.12
71	34	9	9	4	1	0	0	0	0	0	57	225	0.56
81	70	12	2	0	0	0	0	0	0	0	84	491	14.17
82	59	13	5	0	0	0	0	0	0	0	77	486	32.96
83	20	3	4	3	2	1	3	0	0	1	37	147	0.6
85	22	9	6	3	3	0	1	1	0	0	45	252	0.46
91	43	13	7	0	2	0	0	0	0	0	65	488	6.83
92	45	15	2	3	0	0	1	0	0	0	66	449	2.03

Table 1, Distribution of sample pixels among PSUs for the upper Midwest region of the MRLC

The second alternative retains equal inclusion probabilities for pixels within a land-cover stratum. In this option, both the first and second stages of sampling are unequal probability sampling designs, but the combination of the two stages is implemented to achieve equal inclusion probabilities for the overall selection process (*Kish, 1992*). At the first stage, the PSU inclusion probabilities are proportional to the number of pixels of land-cover class k in the PSU. For example, if PSU i has 200 pixels of class k and PSU j has 20 pixels of class k, the probability of sampling PSU i is set at 10 times the probability of selecting PSU j. At the second stage, an equal number of pixels are sampled from each first-stage PSU regardless of the number of pixels in the PSU. For example, suppose two pixels are sampled from each of PSUs i and j. The pixels in PSU i have conditional probability of 2/200 or 0.01 of being sampled at the second stage, and the pixels in PSU j each have conditional probability 2/20 or 0.1 of being selected (the

conditioning is on the PSUs sampled at the first stage). The overall inclusion probability for each pixel is then the product of the conditional second-stage inclusion probability with the first-stage inclusion probability for the PSU. Consequently, the inclusion probabilities in the example would be equal. The high probability of sampling PSU i (relative to PSU j) is compensated for at the second stage in which the pixels in PSU j have 10 times the probability of being sampled relative to the pixels in PSU i. A major advantage of this protocol is that the equal probability feature of sampling in each land-cover class allows for simpler analyses. Accuracy estimates can be obtained from standard stratified random sampling formulas.

Precision is one criterion on which to compare these different two-stage sampling options. Evaluating the relative precision and cost of the design alternatives is difficult because it requires detailed information on the spatial autocorrelation of classification error. That is, relative precision depends on the within-PSU correlation of classification error, and good estimates of this intracluster correlation are generally unavailable. Comparisons based on hypothetical values can provide some insight on the relative precision of the alternatives.

4. Sampling Theory

The sampling theory required to support the two-stage sampling designs described in the previous section is reported by *Sarndal et al. (1992, Sections 9.1-9.4)*. This theory is more complex than that required for simple designs and analyses such as those based on simple random or stratified random sampling. However, the theory is generally applicable and requires only that the inclusion probabilities of the design are known. These inclusion probabilities should be available for any specific choice of first- and second-stage options within the general two-stage structure defined.

The two-stage design currently used for the MRLC assessment creates no novel concerns for estimating the parameters summarizing an error matrix. These estimates follow from standard stratified sampling formulas. The two-stage cluster design does affect standard error calculations because stratified random sampling formulas do not take clustering into account. Standard error approximations ignoring clusters typically underestimate variability (*Stehman, 1997*). Estimating standard errors for cluster sampling requires additional record keeping to identify from which PSU each sample pixel has arisen, and the estimation formulas themselves are more complex. For some accuracy assessment objectives, an approximate standard error may be adequate. Many consumers of the land-cover map will be far more interested in the accuracy estimates themselves than in the standard errors of these estimates. For scientific purposes, a good (i.e., nearly unbiased) estimate of the standard error is important, but for practical use of the accuracy information, an approximate standard error will often suffice.

Another important element of the sampling theory needed for accuracy assessment is subpopulation estimation. A subpopulation (also sometimes called a 'domain') is defined as any subset of the map population for which accuracy estimates are desired. Subpopulations may include geographic regions defined by administrative units (e.g., states or provinces), ecoregions, regions of homogeneous land cover, or subgroups defined by perceived quality of the reference data, such as defined by a confidence rating (*Zhu et al., in press*). Subpopulation estimates can be computed as long as some elements of the subpopulation appear in the sample. Subpopulations not identified as strata may lack adequate sample size for precise estimation if the

subpopulation is small. *Sarndal et al. (1992, Sec. 10.3)* provide the necessary theory for subpopulation estimation.

When the land-cover map has many users and intended applications, accuracy assessment data will be subjected to a variety of user-specified, secondary data analyses. Common examples of secondary analyses are subregional estimates and estimating accuracy for combinations of land-cover classes (e.g., combining low-density residential, high-density residential and high-density commercial into a single class). Simplicity for secondary data analysis is a desirable design feature, and a self-weighting sampling design creates this feature. Equal probability sampling designs are self-weighting (i.e., the sample data need not be weighted in the analysis). Simple random and systematic sampling, both being equal probability designs (*Stehman, 1999*), are examples of self-weighting designs, but these basic designs are rarely cost-effective for a large-area accuracy assessment. More complex self-weighting designs, typically employing cluster sampling, will be necessary. In fact, a fully self-weighting design is often impractical for accuracy assessment. The regional stratification in the MRLC design creates unequal probabilities for sample data arising in different regions, and the common practice of stratifying by land-cover class requires a weighted analysis whenever estimating a parameter combining several land-cover classes. Because a self-weighting design is typically not the best choice for accuracy assessment, the question arises of how to provide publicly accessible reference data. If users conduct secondary analyses without the proper weighting, the estimates will be statistically inconsistent and potentially badly biased. Simply providing a map or database of reference sample locations is inadequate to ensure proper use of the reference data. If reference data are to be made publicly available, a strategy for enhancing proper use of the data must be established.

5. Discussion

Although recent efforts at probability sampling-based accuracy assessments have been successful (*Edwards et al., 1998; Scepan, 1999; Zhu et al., in press*), large-area accuracy assessment based on sound statistical sampling and analysis principles is still in its infancy. Each new experience teaches more about the process, and communicating lessons learned from each effort is important. Sampling design for accuracy assessment of major land-cover mapping projects is challenging because all accuracy objectives cannot be anticipated, and the assessment will be asked to answer far more questions than the resources available could possibly achieve. An additional caution in accuracy assessment planning is that once a sampling design is selected, overcoming the inertia established by the choice can be difficult. Once a design is in place, it will be difficult to revise the design except in minor ways. Typically the time and resources for planning the assessment are severely constrained as the urgency to get a design in place and collect reference data becomes the dominant force. The immediate need to complete the assessment often prevents careful evaluation of different sampling design options, and few resources and little time are allocated for researching better design alternatives. In the MRLC assessment, design choices made very early in the project have by necessity been carried through to subsequent geographic regions. Fortunately, we are in the enviable, but perhaps unusual position of being able to improve the design as we move to new geographic regions.

The MRLC provides a good illustration of how a sampling design created for one set of objectives may have extended influence to subsequent projects. A proposal exists to construct another land-cover map for the conterminous United States based on year 2000 imagery. Initial planning for accuracy assessment of the 2000 map will likely be influenced to some extent by the

design implemented for the 1990 MRLC. Further, the 1990 and 2000 maps create the opportunity for a change-detection product, and significant cost-savings could be achieved if the 1990 reference data can be used as part of the change-detection accuracy assessment. The immediacy of planning the initial regional assessments for the 1990 MRLC precluded considering the potential impact of the design choices on future applications such as change-detection accuracy assessment. It would not necessarily be the case that the 1990 design structures prove advantageous to other uses. However, the general sampling framework implemented for the current MRLC assessment has the necessary flexibility that much of the design structure will be effective for accuracy assessment of the MRLC 2000 land-cover map and change-detection products.

Large-area accuracy assessments still have much room for improvement. We have identified comparing precision for various options within the general two-stage sampling framework as a pressing research need. At a more general level, a clearer framework for interpreting accuracy assessment results in light of applications of the land-cover map is needed. Descriptive accuracy information is important to a broad class of users, but it is not clear that standard accuracy assessment reporting effectively serves the needs of users interested in different spatial scales and/or aggregation of the data. There is also a need to consider the different objectives of designing and reporting accuracy assessment for user consumption versus scientific scrutiny. Map users may be satisfied with a few easy to interpret, approximate accuracy statistics. To satisfy the more demanding task of withstanding scientific scrutiny, statistically rigorous, high quality estimates may be required. Resolving these practical and conceptual issues of large-area accuracy assessments will continue to be a challenging and productive activity.

References

- *Anderson, J.F., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976*, A land use and land cover classification system for use with remote sensor data, U. S. Geological Survey Professional Paper 964, 28 pp.
- *Bellhouse, D.R., 1988*, A brief history of sampling methods, in: Handbook of Statistics, Vol. 6: Sampling, Krishnaiah, P.R., Rao, C.R., Eds., Elsevier Science Publishers, pp. 1-14.
- *Cochran, W.G., 1977*, Sampling Techniques (3rd edition), John Wiley & Sons.
- *Edwards, T.C., Jr., Moisen, G.G., Cutler, D.R., 1998*, Assessing map accuracy in an ecoregion-scale cover-map, Remote Sensing of Environment 63: 73-83.
- *Kish, L., 1987*, Statistical Design for Research, John Wiley & Sons.
- *Kish, L.,1992*, Weighting for unequal Pi. Journal of Official Statistics 8: 183-200.
- *Nusser, S.M., Goebel, J.J., 1997*, The National Resources Inventory: A long-term multi-resource monitoring program, Environmental and Ecological Statistics 4: 181-204.
- *Peterson, S.A., Urquhart, N.S., Welch, E.B., 1999*, Sample representativeness: A must for reliable regional lake condition estimates, Environmental Science and Technology, 33: 1559-1565.
- *Sarndal, C.E., Swensson, B., Wretman, J., 1992*, Model-Assisted Survey Sampling, Springer-Verlag.
- *Scepan, J., 1999*, Thematic validation of high-resolution global land-cover data sets, Photogrammetric Engineering & Remote Sensing 65: 1051-1060.
- *Stehman, S.V., 1997*, Estimating standard errors of accuracy assessment statistics under cluster sampling, Remote Sensing of Environment 60: 258-269.

Stehman, S.V., 1999, Basic probability sampling designs for thematic map accuracy assessment, International Journal of Remote Sensing 20: 2347-2366.

Stehman, S.V., Czaplewski, R.L., 1998, Design and analysis for thematic map accuracy assessment: Fundamental principles, Remote Sensing of Environment 64: 331-344.

USFS, 1992, Forest Service Resource Inventories: An Overview, USGPO 1992-341-350/60861, U.S. Department of Agriculture, Forest Service, Forest Inventory, Economics, and Recreation Research, Washington, DC, 39 pp.

Vogelmann, J.E., Sohl, T.L., Campbell, P.E., Shaw, D.M., 1998, Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources, Environmental Monitoring and Assessment 51: 415-428.

Yang, L., Stehman, S.V., Wickham, J.D., Smith, J.H.,, VanDriel, N.J., 2000, Thematic validation of land cover data of the Eastern United States using aerial photography: Feasibility and challenges, Proceedings of the 4th Spatial Accuracy Conference (these proceedings).

Zhu, Z., Yang, L., Stehman, S.V., Czaplewski, R.L., in press, Accuracy assessment of the USGS regional land cover characterization: New York and New Jersey, Photogrammetric Engineering & Remote Sensing.

Accuracy 2000

Proceedings of
The 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Amsterdam, July 2000

Editors
G.B.M. Heuvelink
M.J.P.M. Lemmens